Recommended Search of Documents from Conversation with Relevant Keywords Using Text Similarity

Keerthana S
Assistant Professor, Computer Science and Engineering, Akshaya College of Engineering and Technology, Coimbatore, India.

Abstract – Data Mining is the procedure of extracting information from huge sets of data. Natural Language Processing is the capability of a computer program that understands the human speech when spoken. Applications of Natural Language Processing are automatic summarization and information retrieval. The keyword extraction and clustering from each short conversation fragment is used to retrieve a small number of relevant documents that can be recommended to participants. Latent Dirichlet Allocation topic modeling technique is used to extract topics and keywords from the Automatic Speech Recognition System. This method is used to derive multiple topically separated queries from the keyword set to maximize the chance of recommendation from the dataset. The word co-occurrence similarity and semantic distance evaluation techniques are introduced to increase the relevant search of keywords to retrieve documents in short duration. The explicit query formulation is introduced to retrieve documents as recommended by users.

Index Terms – Keyword extraction; topic modeling; MALLET; document recommendation; just-in-time retrieval system.

1. INTRODUCTION

In Data mining, clustering and summarization are the characteristics of the descriptive model. Document summarization is used to map data into subsets with associated simple descriptions. Natural Language Processing (NLP) is the process of describing the human and computer interaction. NLP involves automatic summarization and machine translation.

Humans in this modern environment are surrounded by a huge wealth of information that are available as documents, databases and multimedia resources. Due to the rapid development of internet, the volume of information on the internet is increasing exponentially. If the information is available in the form of conversation fragments that can be retrieved from meetings, it can be modeled as implicit queries that are constructed in background from the pronounced words through real time Automatic Speech Recognition (ASR) system.

Keyword extraction is the most important method in the information retrieval research. Keyword extraction is a task that identifies a small set of words or key phrases to describe the meaning of the document. The relevant words that match the keywords from the keyword set can be extracted from the conversation fragments to retrieve recommended documents. The ASR system and Topic Modeling techniques can be used to retrieve keywords from short conversation fragments. The topic modeling technique to extract keywords from the conversation fragment is implemented using MALLET tool kit in java to evaluate the performance of the search of documentation with relevant keywords.

2. RELATED WORK

Just-in-time retrieval systems [1] are used to extract keywords from the implicit queries formulated from the conversational input. The implicit queries are extracted from the words that are spoken by users with the help of the speech recognition systems.

2.1. Query Formulation Methods

The Fixit system [2], one of the first systems for document recommendation is known to be as query-free search system.This system made query free information retrieval and the information relevant to the user is offered without explicit request.

Automatic Content Linking Device (ACLD) [3] is a just-in-time retrieval system used by a small group of people in a meeting. The system prepares implicit queries from words recognized through ASR by constantly listening to the meeting. ACLD prepared a list of information needs of users as a list of keywords simultaneously at regular intervals. The retrieved documents are then recommended to users.

Watson system [4] is used to gather contextual information in the form of text of the document that the user is manipulating to retrieve documents from distributed information repositories.These systems observe user interactions with applications and automatically fulfill the needs using Internet information sources.Forenriching television news with articles from the web, a query freesystem was designed using Term Frequency Inverse Document Frequency (TFIDF) weighting.

2.2. Keyword Extraction Methods

The diverse keyword extraction and clustering method [5] used topic modeling technique to extract keywords. Word frequencies [6] and TFIDF [7] values have been used in earliest techniques to rank words for extraction.

The topic modeling techniques are used to extract keywords in the place of just-in-retrieval systems. The topic modeling technique is used with similarity computing methods [8] to obtain the semantic relations between the words. The semantic relations can be obtained from a manually-constructed thesaurus using topic modeling techniques such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocation (LDA).

PLSA was utilized [9] to represent the topics of a transcribed conversation and ranked the words found in the conversation based on topical similarity. Dirichlet prior distribution can be used to improve the retrieval results of keyword extraction.

3. KEYWORD EXTRACTION USING SIMILARITY METHODS

This section provides the description of keyword extraction for recommended documents using similarity and semantic methods. The steps involved in the recommendation of the documentations with the keywords is shown in the following fig. 1.

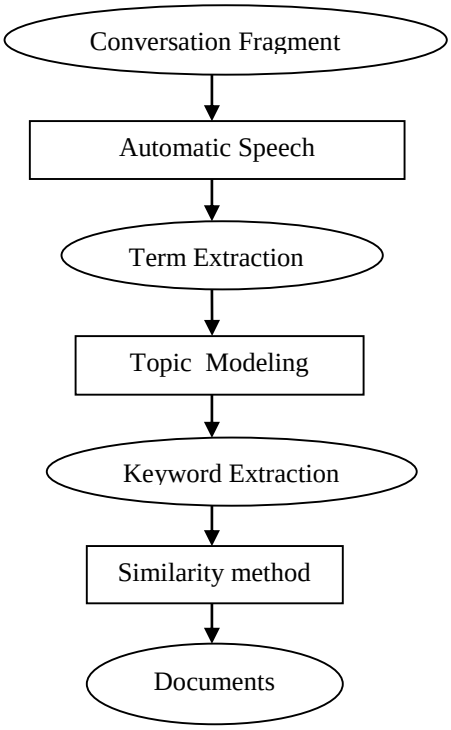

Figure 1 Steps in Document Recommendation

2.1. Automatic Speech Recognition

Automatic Speech Recognition [10] is a technology that permits a system to spot the words that someone speaks into an electro-acoustic transducer or telephone and convert it to transcription. This method begins once a speaker decides to speak a sentence. Then software produces a speech wave kind that embodies the words of the sentence within the spoken input. Then the software decrypts the speech into the simplest estimate of the sentence. It converts the speech signal into a sequence of vectors that are measured throughout the period of the speech signal. Then employing a syntactical decoder it generates a legitimate sequence of representations.

2.2. Term Extraction

The terms of the conversation fragment are extracted from the speech [11]. To extract the important terms from the list of terms, stop words [12] that are referred as verbs, articles are removed. The important terms are extracted for keyword extraction using cosine similarity.

2.3. Keyword Extraction

The extraction of keywords [13] from the transcript of the conversation fragment as provided by the ASR system is by converting to text. The important words are selected as keywords by removing the stop words from the converted text. The advantage of relevant keyword extraction is that the coverage of the most topics of the conversation fragment is maximized. Latent Dirichlet Allocation technique is used as an off-line topic modeling technique [14].

Latent Dirichlet Allocation [15] is a Bayesian probability model composed of word, topic and text. It is a topic modeling and clustering technique that formulates several topically separated queries to build recommended documents. Topic model is used to analyze huge volumes of unlabelled text. A topic is a collection of words that occurs together frequently. Topic model is able to connect words with appropriate meanings and distinguish the words with multi meanings.

LDA is implemented in the MALLET toolkit using java language. The documents are represented as random mixtures over latent topics and each topic is characterized by a distribution over words. Mallet is a Java based package used for natural language processing applications and machine learning applications. It contains efficient, sampling-based implementations of LDA, Pachinko Allocation and Hierarchical LDA [16]. The steps involved in the keyword extraction is shown in the following fig. 2.

The topic modeling technique is used in the extraction of keywords. Initially, a transcript and a topic model that is used to represent the distribution of each abstract topic z for each word w, noted $p(z|w)$ is given as input. The topic model is used to determine the weights for the abstract topics in each conversation fragment, noted β_z.

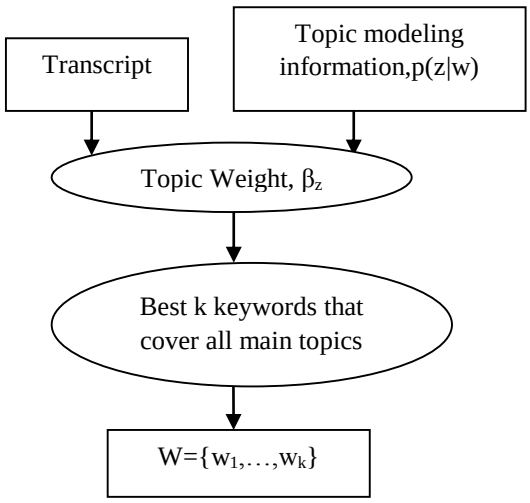

Figure 2 Steps in Keyword Extraction method

$$\beta_z = \frac{1}{N} \sum_{1 \le i \le N} p(z|w_i) \quad (1)$$

Finally the keyword list W = {w1,...,wk} is extracted and it should cover a maximum number of the most important topics in a given fragment. After extracting the keywords using topic modeling similarity methods are applied.

1) COSINE SIMILARITY METHOD

This section provides the description of keyword extraction for recommended documentation using cosine similarity method. The relevant search of documents can be improved by using accurate keywords. The accuracy of keywords and relevant search of documentation can be developed by introducing cosine similarity method. Cosine similarity gives a useful measure of two similar documents that are likely to be in terms of the subject matter.

Cosine similarity is a similarity between two vectors of an inner product space. In information retrieval and text mining, each term is represented as a different dimension and a document is represented by a vector and the value of each dimension is represented by the number of times that term appears in the document. Cosine similarity gives a useful measure of two similar words or documents that are likely to be in terms of the subject matter. The cosine of two vectors can be derived by using

$$P.Q = ||P|| ||Q|| \cos(\theta) \quad (2)$$

Given two vectors of attributes, P and Q, the cosine similarity, cos(θ) is represented using a dot product and magnitude as

$$Similarity = \cos(\theta) = \frac{P.Q}{||P|| * ||Q||} \quad (3)$$

NOTE:

if P = {p1,p2,...,pn} and Q = {q1,q2,...,qn} then

P.B = Sum(p1*q1 + p2*q2 + ... + pn*qn)

||P|| = sqrt(p1^2 + p2^2 + ... + pn^2) and

||Q|| = sqrt(q1^2 + q2^2 + ... + qn^2)

The attribute vectors P and Q are represented as the term frequency vectors of the documents involved in text matching. The cosine similarity is a method of normalizing document length during comparison. For information retrieval, the range of cosine similarity between two documents is from 0 to 1 because the term frequencies cannot be negative. The steps for calculating cosine similarity between two texts:

1. Identify all distinct words in both texts.
2. Identify the frequency of occurrences of these words in both text and treat it as vector.
3. Apply cosine similarity function.

Example: To calculate the cosine similarity between two texts:

Text 1: Julie loves me more than Linda loves me

Text 2: Jana likes me more than Julie loves me

For the given texts mentioned in the above example, the vector representation using frequency of both the texts is mentioned in the following table.

Distinct Words from both texts	Frequency in Text-1	Frequency in Text-2
Julie	1	1
Loves	2	1
Me	2	2
More	1	1
Than	1	1
Linda	1	0
Jana	0	1
Likes	0	1

Table 1 Vector Representation of Texts

Vector A = [1,2,2,1,1,1,0,0]

Vector B = [1,1,2,1,1,0,1,1]

2) WORD CO-OCCURRENCE METHOD

This section describes the improved text similarity computing based on word co-occurrence. The steps involved in the word co-occurrence method is given below.

Step 1: Extract the feature word from the terms extracted from the topics.

Step 2: Calculate the co-occurrence probability of the text feature word.

Assume T_i is the topic of text D_i, word set $W = \{w_{i1}, w_{i2},...,w_{iN}\}$ is the feature word of topic T_i, the co-occurrence probability of the feature word is p_{11}, p_{12}, p_{13},...,P_{NN}. The co-occurrence probability of the text feature word is calculated using

$$(4) \qquad p(w_{im}, w_{in}) = p(w_{im}|T_i)\, p(w_{in}|T_i)$$

Step 3: Calculate the correlation of the arbitrary feature words.

Assume if the probability of feature word w_{im} in topic T_i is p_{im}, then the co-occurrence probability of feature word w_{im} and w_{jn} in topic T_i is p_{mn}. The correlation of the arbitrary feature words is calculated using

$$(5) \qquad correlation(w_{im}, w_{jn}) = \frac{p_{mn}}{p_{im} + p_{jn} - p_{mn}}$$

According to formula, if the value of p_{mn} is 0, correlation $(w_{im},w_{jn}) = 0$, that means feature word w_{im} and $w_{jn}$is uncorrelated. If correlation $(w_{im},w_{jn}) \neq 0$, feature word w_{im} and $w_{jn}$is correlated.

Step 4: Calculate the similarity between the arbitrary texts.

Assume d_i and d_j are arbitrary texts from the document set, V denotes the number of feature word of the selected document. λ€[0,1] denotes the correlation coefficient assigned to the document. If the value of Similarity (d_i,d_j) is smaller, then the two texts d_i and d_j will be more similar. The similarity between the texts is calculated using

$$(6) \quad Similarity(d_i, d_j) = \lambda D_{js}(d_i, d_j) + (1 - \lambda)\left[\sum_{m,n=1}^{v} (1 - correlation(w_{im}, w_{jn}))\Big/(V(V-1))\right]$$

3) SEMANTIC BASED COMPUTATION

The structured network with semantic distance between the words is used to measure the importance of words.The network structure not only uses the matching of synonym and near-synonym but also the synonyms dictionary that increases the accuracy of semantic calculation.The synonyms dictionary has a detailed classification and every word has corresponding code.

$$(7) \qquad Code_i = X_{i1}X_{i2}X_{i3}X_{i4}X_{i5}F_i$$

The semantic distance $Dis(w_1,w_2)$ between words w_1 and w_2 is defined as

$$(8) \qquad Dis(w_1,w_2) = \min_{i=1,2,\ldots m;j=1,2,\ldots n} Dis(code_{1i},code_{2j})$$

Thus the relevant keywords can be extracted using the similarity methods for retrieving recommended documents.

2.4. Keyword Clustering

A cluster of keywords are formed by ranking keywords [17] for each main topic of the fragment. Keywords with high value of p(z|w) will be ranked higher in the cluster of topic z and these keywords will be selected from the topics with high value of β_z.Then clusters are ranked based on their β_z values.

2.5. Document Recommendation

Implicit queries can be prepared for each conversation fragment by using all keywords selected by the keyword extraction technique. The retrieval results can be improved by formulating multiple implicit queries for each conversation fragment with the keywords of each cluster. Explicit queries can also be formulated to retrieve documents as recommended by users. The similarity methods result better in the retrieval of documents with relevant keywords.

4. RESULTS AND DISCUSSIONS

The technique for relevant keyword extraction for recommended documentation is experimented over a sample conversation fragment from ELEA corpus.The topic modeling technique will retrieve a list of terms from the conversation fragment and the keywords are extracted from the terms. The similarity methods and explicit query request will result in the retrieval of relevant keywords and documents. The sample conversation fragment that is converted into text is shown in fig. 2.

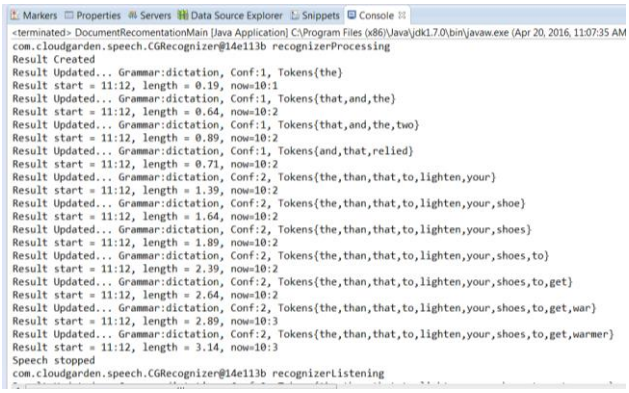

Figure 3 Conversation Fragment in the text format

The term extraction result is shown in fig. 3. The keyword extraction result is shown in fig. 4. The documents for the extracted keywords result is shown in fig. 5. The explicit query given by the user is shown in the fig. 6. The documents for the given explicit query is shown in the fig. 7. The comparison of keyword extraction methods is shown in the fig. 8.

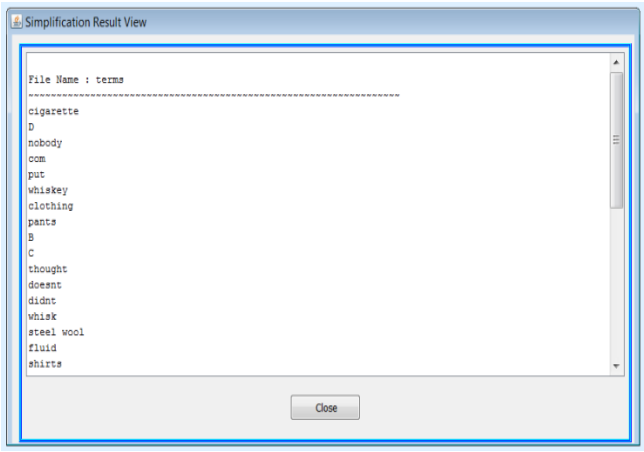

Figure 4 Terms Extraction

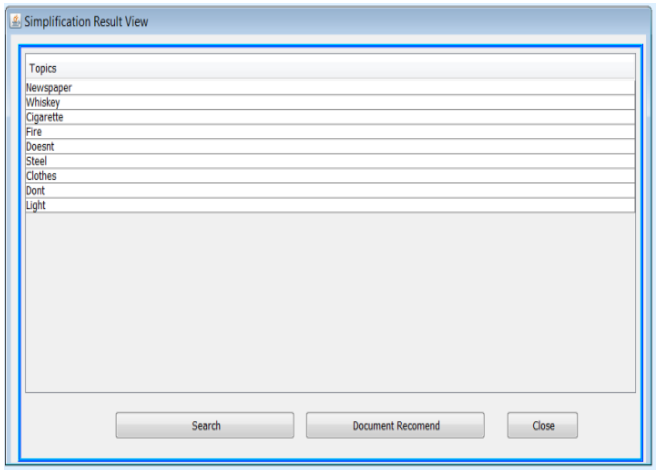

Figure 5 Keyword Extraction

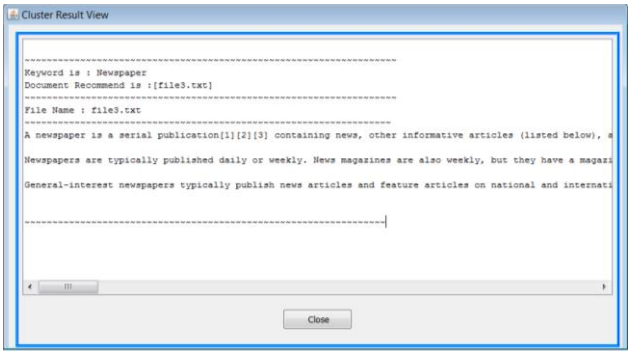

Figure 6 Recommended Documentation

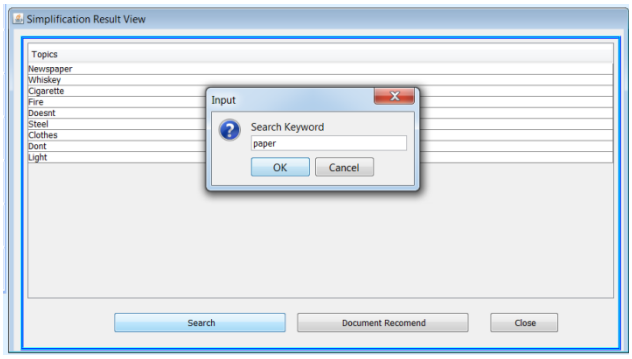

Figure 7 Explicit Query

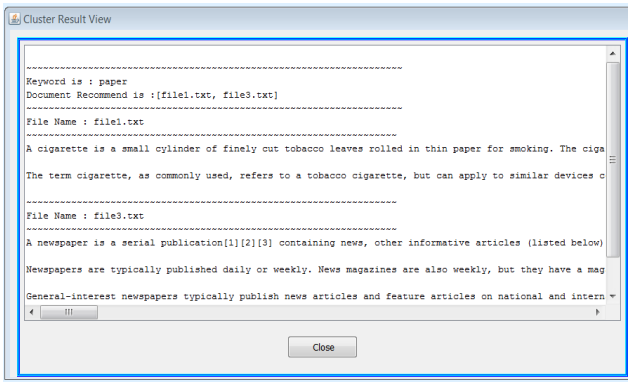

Figure 8 Documents for Explicit Query

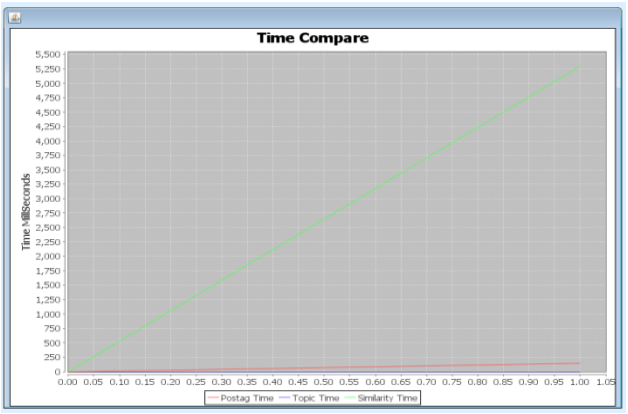

Figure 9 Comparison of Keyword Extraction

The Keyword Extraction technique with cosine similarity is compared with the keyword extraction technique with word co-occurrence and semantic distance. The Keyword Extraction technique with word co-occurrence, semantic distance and explicit query request methods extract more relevant keywords than the previous technique. The explicit query request is also introduced to retrieve documents recommended by the user. The examples of keyword sets obtained by the keyword extraction methods for a fragment of the ELEA corpus is shown in the following table.

Keyword Extraction with Cosine Similarity	Keyword Extraction with word co-occurrence and semantic distance
S={newspaper, whiskey, cigarette, fire, doesnt matter, steel wool, clothes, light}	S={newspaper, whiskey, cigarette, fire, dont, steel, clothes, light}

Table 2 Examples Of Keyword Sets Obtained By The Keyword Extraction Methods For A Fragment Of The Elea Corpus

5. CONCLUSION

The relevant search of documentation with accurate keywords is focused on retrieving more relevant documents recommended by the users. It is focused on modeling the information needs of the users by deriving implicit and explicit queries from short conversation fragments. These queries are based on sets of keywords extracted from the conversation. The LDA topic modeling technique is used to extract keywords based on the more relevant topics. A clustering technique is used to divide the set of keywords into smaller topically independent subsets constituting implicit queries. The similarity methods introduced in the search of documents extracted relevant keywords to retrieve more relevant documentations recommended by the users.

REFERENCES

[1] A. Popescu-Belis, M. Yazdani, A. Nanchen and P. N. Garner (2011), 'A speech-based just-in-time retrieval system using semantic search', in Proceedings Annual Conference (HLT-NAACL), pp. 80–85.

[2] P. E. Hart and J. Graham (1997), 'Query-free information retrieval', International Intelligence System Technology Application, vol. 12, no.5, pp.32–37.

[3] D. Sanchez-Cortes, O. Aran, M. Schmid Mast and D. Gatica-Perez (2013), 'A nonverbal behavior approach to identify emergent leaders in small groups', IEEE Transaction Multimedia, vol. 14, no. 3, pp. 816–832.

[4] J. Budzik and K. J. Hammond (2000), 'User interactions with everyday applications as context for just-in-time information access', in Proceedings 5thInternational Conference Intelligence User Interfaces (IUI'00), pp. 44–51.

[5] Maryam Habibi and Andrei Popescu-Belis (2015), 'Keyword Extraction and Clustering for Document Recommendation in Conversations', IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 23, no. 4.

[6] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM J. Res. Develop.*, vol. 1, no.4, pp. 309–317, 1957.

[7] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage. J., vol. 24, no. 5, pp. 513–523,1988.

[8] Minglai Shao (2014), 'Text Similarity Computing based on LDA Topic Model and Word Co-occurrence' in International Conference on Software Engineering, Knowledge Engineering and Information Engineering.

[9] D. Harwath and T. J. Hazen, "Topic identification based extrinsic evaluationof summarization techniques applied to conversational speech," in Proceedings International Conference Acoustic, Speech, Signal Process. (ICASSP), 2012,pp. 5073–5076.

[10] C. Cieri, D. Miller and K. Walker (2004), 'The Fisher Corpus: A resource for the next generations of speech-to-text', in Proc. 4th International Conference Language Resources Evaluation (LREC), pp. 69–71.

[11] Z. Liu, P. Li, Y. Zheng and M. Sun (2009), 'Clustering to find exemplar terms for keyphrase extraction', in Proceedings Conference Empirical Method Natural Language Processing (EMNLP'09), pp. 257–266.

[12] M. Habibi and A. Popescu-Belis (2012), 'Using crowdsourcing to compare document recommendation strategies for conversations', Workshop Recommendation Utility Evaluation: Beyond RMSE (RUE'11), pp. 15–20.

[13] Hulth (2003), 'Improved automatic keyword extraction given more linguistic knowledge', in Proceedings Conference Empirical Methods Natural Language Processing (EMNLP'03), pp. 216–223.

[14] Z. Liu, W. Huang, Y. Zheng and Sun M. (2010), 'Automatic keyphrase extraction via topic decomposition', in Proceedings Conference Empirical Method Natural Language Processing (EMNLP'10), pp. 366–376.

[15] M. Habibi and A. Popescu-Belis (2013), 'Diverse keyword extraction from conversation in Proceedings 51st Annual Meeting Association Computation Linguist., pp. 651–657.

[16] K. Riedhammer, B. Favre and D. Hakkani-Tur (2008), 'A keyphrase based approach to interactive meeting summarization', in Proc. IEEE Spoken Language Technology Workshop (SLT'08), pp. 153–156.

[17] Y. Matsuo and M. Ishizuka (2004), 'Keyword extraction from a single document using word co-occurrence statistical information', International J. Artificial Intelligence Tools, vol. 13, no. 1, pp. 157–169.